

Domain adaptation based Speaker Recognition on Short Utterances

Ahilan Kanagasundaram^{*,+}, David Dean^{*}, Sridha Sridharan^{*} and Clinton Fookes^{*}

Speech and Audio Research Laboratory^{*}
Queensland University of Technology, Brisbane, Australia^{*}
{a.kanagasundaram, d.dean, s.sridharan, c.fookes}@qut.edu.au^{*}
Electrical & Electronic Engineering, Faculty of Engineering⁺
University of Jaffna, Jaffna, Sri Lanka⁺
ahilan@eng.jfn.ac.lk⁺

Abstract

This paper explores how the in- and out-domain probabilistic linear discriminant analysis (PLDA) speaker verification behave when enrolment and verification lengths are reduced. Experiment studies have found that when full-length utterance is used for evaluation, in-domain PLDA approach shows more than 28% improvement in EER and DCF values over out-domain PLDA approach and when short utterances are used for evaluation, the performance gain of in-domain speaker verification reduces at an increasing rate. Novel modified inter dataset variability (IDV) compensation is used to compensate the mismatch between in- and out-domain data and IDV-compensated out-domain PLDA shows respectively 26% and 14% improvement over out-domain PLDA speaker verification when SWB and NIST data are respectively used for S normalization. When the evaluation utterance length is reduced, the performance gain by IDV also reduces as short utterance evaluation data i-vectors have more variations due to phonetic variations when compared to the dataset mismatch between in- and out-domain data.

Index Terms: speaker verification, GPLDA, IDV, i-vectors, domain adaptation

1. Introduction

In a typical speaker verification system, the significant amount of speech is required for speaker model enrolment and verification, especially in the presence of large intersession variability. Techniques based on factor analysis, such as joint factor analysis (JFA) [1, 2], i-vectors [3] and probabilistic linear discriminant analysis (PLDA) [4], have demonstrated outstanding behaviour in National Institute of Standards and Technology (NIST) conditions [5, 6]. Unfortunately, the performance of many of these approaches degrades rapidly as the available amount of enrolment and/or verification speech decreases [7, 8, 9], limiting the utility of speaker verification in real world applications, such as access control or forensics.

Recent studies have found that when speaker verification is developed on data from the Switchboard database and evaluated using data from NIST evaluations, the dataset mismatch significantly affects the speaker verification performance [10, 11, 12]. Several techniques have been proposed to address this issue and achieve state-of-the-art speaker verification performance when speaker verification developed in one domain data and evaluated in different domain data.

The main aim of this paper is to investigate the effect of only having short utterances available for evaluation for the out- and in-domain PLDA speaker verification. Previous studies

have shown that when PLDA speaker verification is developed using out-domain data and evaluated using in-domain long utterances, it fails to achieve good performance due to mismatch between out- and in-domain data. In this paper, it is analysed whether the above outcome is true with short utterance evaluation data as well. Subsequently, novel modified inter dataset variability (IDV) compensation approach is studied with short utterance evaluation data in order to find whether those approaches improve the performance of out-domain short utterance PLDA speaker verification system. We also analyse the speaker verification performance with regards to the duration of utterances used for both speaker evaluation (enrolment and verification) and score normalization.

This paper is structured as follows: Section 2 details the i-vector feature extraction techniques. Section 3 details the inter dataset variability compensation approach. Section 4 explains the Gaussian PLDA (GPLDA) based speaker verification system. The experimental protocol and corresponding results are given in Section 5 and Section 6. Section 7 concludes the paper.

2. I-vector feature extraction

I-vectors represent the Gaussian mixture model (GMM) super-vector by a single total-variability subspace. This single-subspace approach was motivated by the discovery that the channel space of JFA contains information that can be used to distinguish between speakers [13]. An i-vector speaker and channel dependent GMM super-vector can be represented by,

$$\mu = \mathbf{m} + \mathbf{T}\mathbf{w}, \quad (1)$$

where \mathbf{m} is the same universal background model (UBM) super-vector used in the JFA approach and \mathbf{T} is a low rank total-variability matrix. The total-variability factors (\mathbf{w}) are the i-vectors, and are normally distributed with parameters $N(0, I)$. Extracting an i-vector from the total-variability subspace is essentially a *maximum a-posteriori adaptation* (MAP) of \mathbf{w} in the subspace defined by \mathbf{T} . An efficient procedure for the optimization of the total-variability subspace \mathbf{T} and subsequent extraction of i-vectors is described Dehak *et al.* [3, 14]. In this paper, the pooled total-variability approach is used for i-vector feature extraction where the total-variability subspace ($R_w^{telmic} = 500$) is trained on SWB dataset.

3. Modified IDV compensation approach

When PLDA speaker verification is trained using out-domain data, it significantly affects the speaker verification performance

due to the mismatch between evaluation and development data. Recently, we proposed inter dataset variability (IDV) compensation approach to compensate the mismatch between out- and in-domain data [15]. In this paper, we propose modified IDV compensation approach to effectively compensate the mismatch between out- and in-domain data. In previous IDA approach, dataset mismatch variation was calculated using the outer product of the difference between the out-domain i-vectors and average of speaker unlabelled in-domain i-vectors. In this new approach, dataset mismatch between in- and out-domain data is estimated using outer product of difference between the out-domain i-vectors and average of speaker unlabelled in-domain i-vectors, and outer product of difference between in-domain i-vectors and average of speaker unlabelled out-domain i-vectors. The dataset mismatch variation, \mathbf{S}'_{IDV} , can be calculated as follows,

$$\mathbf{S}'_{IDV} = \frac{1}{N_1} \sum_{i=1}^{N_1} (\mathbf{w}_i^{OD} - \bar{\mathbf{w}}^{ID})(\mathbf{w}_i^{OD} - \bar{\mathbf{w}}^{ID})^T + \frac{1}{N_2} \sum_{j=1}^{N_2} (\mathbf{w}_j^{ID} - \bar{\mathbf{w}}^{OD})(\mathbf{w}_j^{ID} - \bar{\mathbf{w}}^{OD})^T \quad (2)$$

where \mathbf{w}_i^{OD} and \mathbf{w}_j^{ID} are out- and in-domain i-vectors. $\bar{\mathbf{w}}^{OD}$ and $\bar{\mathbf{w}}^{ID}$ are mean of out- and in-domain i-vectors. N_1 and N_2 are number of out- and in-domain i-vectors. SWB data was chosen as out-domain data and 150 speaker with 10 sessions were randomly selected from NIST 2004, 2005 and 2005 telephone data as in-domain data. The decorrelated matrix, \mathbf{D}' , is calculated using the Cholesky decomposition of $\mathbf{D} \mathbf{D}'^T = \frac{1}{\mathbf{S}'_{IDV}}$. Inter dataset variability compensated out-domain i-vectors are extracted as follows,

$$\mathbf{w}_{modified-IDV} = \mathbf{D}'^T \mathbf{w} \quad (3)$$

Once inter-dataset variability compensated i-vectors, LDA projection is applied to compensate the additional session variation prior to the PLDA modelling and reduce the dimensionality [16], which is explained in following in Section 3.1.

3.1. LDA approach

The LDA transformation is estimated based up the standard within- and between-class scatter estimations \mathbf{S}_b and \mathbf{S}_w , calculated as

$$\mathbf{S}_b = \sum_{s=1}^S n_s (\bar{\mathbf{w}}_s - \bar{\mathbf{w}})(\bar{\mathbf{w}}_s - \bar{\mathbf{w}})^T, \quad (4)$$

$$\mathbf{S}_w = \sum_{s=1}^S \sum_{i=1}^{n_s} (\mathbf{w}_i^s - \bar{\mathbf{w}}_s)(\mathbf{w}_i^s - \bar{\mathbf{w}}_s)^T, \quad (5)$$

where S is the total number of speakers, n_s is number of utterances of speaker s . The mean i-vectors, $\bar{\mathbf{w}}_s$ for each speaker, and $\bar{\mathbf{w}}$ is the across all speakers are defined by

$$\bar{\mathbf{w}}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbf{w}_i^s, \quad (6)$$

$$\bar{\mathbf{w}} = \frac{1}{N} \sum_{s=1}^S \sum_{i=1}^{n_s} \mathbf{w}_i^s. \quad (7)$$

where N is the total number of sessions. In the first stage, LDA attempts to find a reduced set of axes \mathbf{A} through the eigenvalue

decomposition of $\mathbf{S}_b \mathbf{v} = \lambda \mathbf{S}_w \mathbf{v}$. The IDV-compensated LDA-projected i-vector can be calculated as follows,

$$\hat{\mathbf{w}}_{IDV-LDA} = \mathbf{A}^T \mathbf{w} \quad (8)$$

After LDA-projection, length-normalized GPLDA model parameters are estimated in as described in Section 4.

4. Length-normalized GPLDA system

4.1. PLDA modelling

In this paper, we have chosen length-normalized GPLDA, as it is also a simple and computationally efficient approach [17]. The length-normalization approach is detailed by Garcia-Romero *et al.* [17], and this approach is applied on development and evaluation data prior to GPLDA modelling. A speaker and channel dependent length-normalized i-vector, $\hat{\mathbf{w}}_r$ can be defined as,

$$\hat{\mathbf{w}}_r = \bar{\mathbf{w}} + \mathbf{U}_1 \mathbf{x}_1 + \varepsilon_r \quad (9)$$

where for given speaker recordings $r = 1, \dots, R$; \mathbf{U}_1 is the eigenvoice matrix, \mathbf{x}_1 is the speaker factors and ε_r is the residuals. In the PLDA modeling, the speaker specific part can be represented as $\bar{\mathbf{w}} + \mathbf{U}_1 \mathbf{x}_1$, which represents the between speaker variability. The covariance matrix of the speaker part is $\mathbf{U}_1 \mathbf{U}_1^T$. The channel specific part is represented as ε_r , which describes the within speaker variability. The covariance matrix of channel part is Λ^{-1} . We assume that precision matrix (Λ) is full rank. Prior to GPLDA modelling, standard LDA approach is applied to compensate the additional channel variations as well as reduce the computational time [18].

4.2. GPLDA scoring

Scoring in GPLDA speaker verification systems is conducted using the batch likelihood ratio between a target and test i-vector [4]. Given two i-vectors, \mathbf{w}_{target} and \mathbf{w}_{test} , the batch likelihood ratio can be calculated as follows,

$$\ln \frac{P(\mathbf{w}_{target}, \mathbf{w}_{test} | H_1)}{P(\mathbf{w}_{target} | H_0)P(\mathbf{w}_{test} | H_0)} \quad (10)$$

where H_1 denotes the hypothesis that the i-vectors represent the same speakers and H_0 denotes the hypothesis that they do not.

5. Experimental methodology

The proposed methods were evaluated using the the NIST 2008 SRE corpora. The shortened evaluation utterances were obtained by truncating the NIST 2008 *short2-short3* condition to the specified length of active speech for both enrolment and verification. Prior to truncation, the first 20 seconds of active speech were removed from all utterances to avoid capturing similar data across multiple utterances. For NIST 2008, the performance was evaluated using the equal error rate (EER) and the minimum decision cost function (DCF), calculated using $C_{miss} = 10$, $C_{FA} = 1$, and $P_{target} = 0.01$ [5]. Outer-domain data is defined as Switchboard I, II phase I, II, III corpora, and in-domain data is defined as NIST 2004, 2005 and 2006 SRE corpora.

We have used 13 feature-warped MFCC with appended delta coefficients and two gender-dependent UBMs containing 512 Gaussian mixtures throughout our experiments. The UBMs were trained on out-domain data, and then used to calculate the Baum-Welch statistics before training a gender dependent

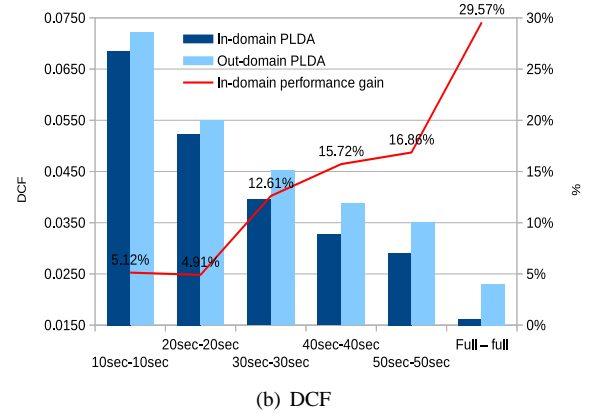
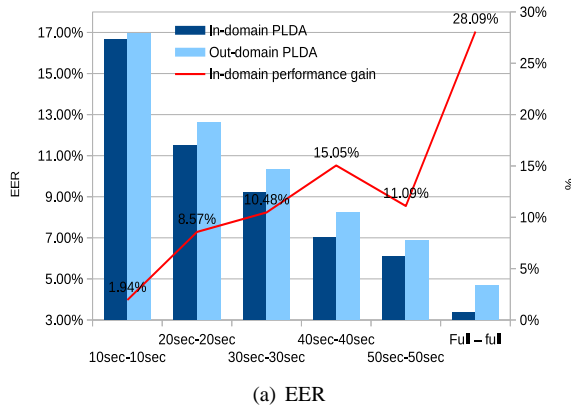


Figure 1: The performance comparison of in- and out-domain PLDA speaker verification on common set of NIST 2008 different short utterance evaluation conditions is shown in bar chart. The performance gain of in-domain PLDA over out-domain PLDA is also shown in line graph (a) EER, (b) DCF.

Table 1: Performance comparison of IDV-compensated and modified IDV-compensated PLDA systems on NIST 2008 short2-short3 evaluation condition. The best performing systems by both EER are highlighted across each row.

System	Without Snorm	With Snorm
Out-domain PLDA	4.86%	3.85%
IDV-compensated PLDA	4.37%	3.55%
Modified IDV-compensated PLDA	3.79%	3.29%

Table 2: Performance comparison of modified IDV-compensated PLDA system with full and matched length score normalization data. The best performing systems by both EER are highlighted across each row.

Evaluation utterance	S-norm development data	
	Full-length	Matched length
10sec - 10 sec	17.63%	17.64%
20sec - 20 sec	12.36%	12.36%
30sec - 30 sec	9.47%	9.47%
40sec - 40 sec	7.41%	7.09%
50sec - 50 sec	6.09%	5.85%

total-variability subspace of dimension $R_w = 500$. The pooled total-variability representation was trained using out-domain data. For out-domain PLDA speaker verification system, the GPLDA parameters were trained using Switchboard I, II phase I, II, III corpora. We empirically selected the number of eigen-voices (N_1) equal to 120 as best value according to speaker verification performance over an evaluation set. 150 eigenvectors were selected for LDA estimation. S-normalisation was applied for experiments. 150 NIST telephone speakers with 10 sessions were used for NIST S-normalization data, and randomly selected utterances from Switchboard I, II phase I, II, III were pooled to form the Switchboard S-normalisation dataset [19].

6. Results and Discussions

6.1. Analysis of in- and out-domain PLDA speaker verification

Previous studies have found that when speaker verification is developed in one domain and evaluated in a different domain, it significantly affects the speaker verification performance [10]. In this section, in- and out-domain PLDA speaker verification systems are closely studied with short utterance evaluation data.

Figure 1 depicts results comparing the performance of in- and out-domain PLDA speaker verification on short utterance evaluation conditions. It can be observed from the EER performance indicated by the bars that in-domain PLDA approach shows better performance than out-domain PLDA approach in

all short utterance evaluation conditions. However, it was also observed from the in-domain performance gain indicated by the red line in Figure 1 that when full-length utterance is used for evaluation, in-domain PLDA approach shows more than 28% improvement in EER and DCF values over out-domain PLDA approach and when evaluation utterance length reduces the performance gain of in-domain PLDA approach reduces in increasing rate. Further, when very short utterance are used for evaluation, there is no performance difference between in- and out-domain PLDA speaker verification as short utterance evaluation data i-vectors have more variations due to phonetic variations when compared to the dataset mismatch between in- and out-domain data. Results also suggest that data mismatch between in- and out-domain data is not influencing the performance of short utterance-based PLDA speaker verification system.

6.2. Analysis of modified IDV-compensated out-domain PLDA speaker verification

In previous section, we have found that when speaker verification is evaluated on very short utterances, dataset mismatch doesn't influence the performance as short utterance evaluation data has a lot of uncertainties due to phonetic variation. In this section, initially both IDV and modified IDV approaches will be analysed on NIST standard conditions and subsequently the best IDV approach will be investigated with short utterance

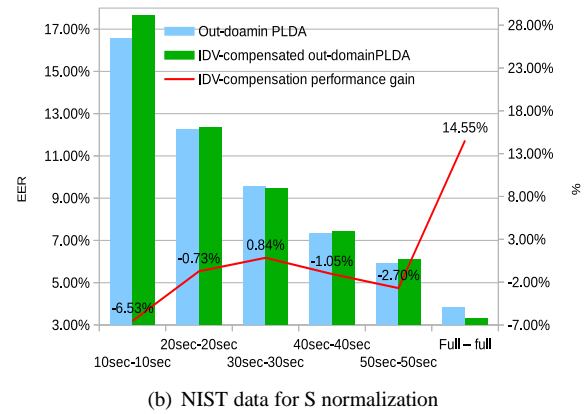
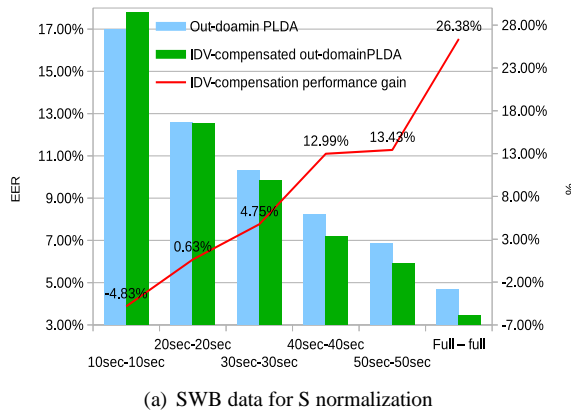


Figure 2: Comparison of the performance of modified IDV-compensated out-domain PLDA speaker verification against out-domain PLDA speaker verification on common set of NIST 2008 different short utterance evaluation conditions is shown in bar chart. The performance gain of modified IDV-compensated PLDA over out-domain PLDA is also shown in line graph (a) SWB data for S normalization, (b) NIST data for S normalization.

evaluation condition.

The Table 1 compares the performance of IDV-compensated and modified IDV-compensated PLDA systems against out-domain PLDA system on NIST 2008 short2-short3 evaluation conditions and it can be observed that modified IDV-compensated PLDA achieves 7% improvement over IDV-compensated PLDA systems.

The performance comparison between modified IDV-compensated out-domain PLDA approach and out-domain PLDA approach is shown in Figure 2. It has been observed with the aid of Figure 2 (a) and (b) that modified IDV-compensated PLDA shows respectively 26% and 14% improvement over out-domain PLDA speaker verification when SWB and NIST data are respectively used for S normalization. It was also observed with the aid of Figure 2 (a) that when evaluation utterance length reduces, the performance gain by modified IDV-compensation also reduces. We believe that this is due to the fact that short utterance i-vectors have more variation arising from the phonetic content compared to variations arising from mismatch between in and out domain data which we compensate using our modified IDV.

6.3. Analysis of PLDA speaker verification with matched length score normalization data

Table 2 presents the results comparing the performance of the modified IDV-compensated PLDA system with full-length score normalization and matched-length score normalization (score normalization data truncated to same length as evaluation data). We found that matched-length score normalization improves the EER performance of modified IDV-compensated PLDA system when utterance length increases above 40 sec. This shows that rather than being a hindrance to normalisation performance, limited development data (if matched in length), can improve normalisation for speaker verification.

7. Conclusion

This paper explored how the in- and out-domain PLDA speaker verification behave when enrolment and verification lengths are

reduced. Our experiment studies have found that when full-length utterance were used for evaluation, in-domain PLDA approach showed more than 28% improvement in EER and DCF values over out-domain PLDA approach and when short utterances were used for evaluation, the performance gain of in-domain speaker verification reduced at an increasing rate. The IDV and modified IDV compensations were used to compensate the mismatch between in- and out-domain data, and found that modified IDV is better approach than IDV approach. Modified IDV-compensated out-domain PLDA showed respectively 26% and 14% improvement over out-domain PLDA speaker verification when SWB and NIST data were respectively used for S normalization. Further when evaluation utterance length reduced, the performance gain by IDV also reduced as short utterance evaluation data i-vectors have more variations due to phonetic variations compared dataset mismatch between in- and out-domain data.

8. Acknowledgements

This project was supported by an Australian Research Council (ARC) Linkage grant LP130100110.

9. References

- [1] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," tech. rep., CRIM, 2005.
- [2] S. Vogt, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.
- [3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2010.
- [4] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic*, 2010.

- [5] “The NIST year 2008 speaker recognition evaluation plan,” tech. rep., NIST, 2008.
- [6] “The NIST year 2010 speaker recognition evaluation plan,” tech. rep., NIST, 2010.
- [7] R. Vogt, B. Baker, and S. Sridharan, “Factor analysis subspace estimation for speaker verification with short utterances,” in *Interspeech 2008*, (Brisbane, Australia), September 2008.
- [8] A. Kanagasundaram, R. Vogt, B. Dean, S. Sridharan, and M. Mason, “i-vector based speaker recognition on short utterances,” in *Proceed. of INTERSPEECH*, pp. 2341–2344, International Speech Communication Association (ISCA), 2011.
- [9] A. Kanagasundaram, D. Dean, S. Sridharan, and R. Vogt, “PLDA based speaker recognition with weighted LDA techniques,” in *Proc. Odyssey Workshop*, 2012.
- [10] D. Garcia-Romero and A. McCree, “Supervised domain adaptation for i-vector based speaker recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 4047–4051, IEEE, 2014.
- [11] H. Aronowitz, “Inter dataset variability compensation for speaker recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 4002–4006, IEEE, 2014.
- [12] O. Glembek, J. Ma, P. Matejka, B. Zhang, O. Pl-chot, L. Burget, and S. Matsoukas, “Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 4032–4036, IEEE, 2014.
- [13] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, “Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification,” in *Proceedings of Interspeech*, p. 1559 1562, 2009.
- [14] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.
- [15] A. Kanagasundaram, D. Dean, and S. Sridharan, “Improving out-domain PLDA speaker verification using unsupervised inter-dataset variability compensation approach,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2015.
- [16] A. Kanagasundaram, D. Dean, S. Sridharan, M. McLaren, and R. Vogt, “I-vector based speaker recognition using advanced channel compensation techniques,” in *Computer Speech and Language*, 2013.
- [17] D. Garcia-Romero and C. Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems,” in *International Conference on Speech Communication and Technology*, pp. 249–252, 2011.
- [18] A. Kanagasundaram, R. Vogt, D. Dean, and S. Sridharan, “PLDA based speaker recognition on short utterances,” in *The Speaker and Language Recognition Workshop (Odyssey 2012)*, ISCA, 2012.
- [19] S. Shum, N. Dehak, R. Dehak, and J. Glass, “Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification,” *Proc. Odyssey*, 2010.